Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230060

Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning

Berk OZTURK[a,1], Tom LAWTON[a,b], Stephen SMITH[a], Ibrahim HABLI[a]
[a] *University of York, York, YO10 5GH, UK*
[b] *Bradford Teaching Hospitals NHS Foundation Trust, Bradford BD9 6RJ, UK*

Abstract. Type 2 diabetes is a life-long health condition, and as it progresses, a range of comorbidities can develop. The prevalence of diabetes has increased gradually, and it is expected that 642 million adults will be living with diabetes by 2040. Early and proper interventions for managing diabetes-related comorbidities are important. In this study, we propose a Machine Learning (ML) model for predicting the risk of developing hypertension for patients who already have Type 2 diabetes. We used the Connected Bradford dataset, consisting of 1.4 million patients, as our main dataset for data analysis and model building. As a result of data analysis, we found that hypertension is the most frequent observation among patients having Type 2 diabetes. Since hypertension is very important to predict clinically poor outcomes such as risk of heart, brain, kidney, and other diseases, it is crucial to make early and accurate predictions of the risk of having hypertension for Type 2 diabetic patients. We used Naïve Bayes (NB), Neural Network (NN), Random Forest (RF), and Support Vector Machine (SVM) to train our model. Then we ensembled these models to see the potential performance improvement. The ensemble method gave the best classification performance values of accuracy and kappa values of 0.9525 and 0.2183, respectively. We concluded that predicting the risk of developing hypertension for Type 2 diabetic patients using ML provides a promising stepping stone for preventing the Type 2 diabetes progression.

Keywords. Type 2 diabetes, comorbidity, machine learning, healthcare, data quality

1. Introduction

Diabetes is one of the most common health conditions in the world, and its incidence in the population has been increasing rapidly over the years [1]. It is a lifelong health condition that occurs when the pancreas cannot produce enough insulin to balance blood sugar (blood glucose) [2]. This health condition can be observed at any age and, if not managed properly, can progress and develop comorbidities [3]. Progression of these comorbidities could result in a range of poor outcomes for the patient, including blindness, the need for dialysis, heart attacks and strokes, the need for limb amputation, and even mortality [4].

[1] Corresponding Author: Berk Ozturk, Tel.: +44 751 337 5115.
E-mail addresses: berk.ozturk@york.ac.uk (B. Ozturk), tom.lawton@bthft.nhs.uk (T. Lawton), stephen.smith@york.ac.uk (S. Smith), ibrahim.habli@york.ac.uk (I. Habli).

Early diagnosis of diabetes progression and proper management of diabetes can significantly prevent the progression of diabetes [5]. Further, it is well known that hypertension is one of the most important risk factors for developing comorbidities [6]. Because diabetes and hypertension are synergistically dangerous, early and accurate prediction of hypertension risk is crucial in terms of managing the progression of diabetes.

Type 1 and Type 2 diabetes are the two main types of diabetes in the world. Type 1 diabetes occurs when the body's immune system targets and destroys the insulin-producing cells in the pancreas [7]. This can develop quickly and requires regular insulin injections [3]. Type 2 diabetes develops when pancreatic cells are not able to produce enough insulin or when body cells do not react to insulin [7]. People may have Type 2 diabetes without realising it because the early symptoms tend to be ambiguous [4]. However, this type of diabetes can be managed by changing lifestyle and employing appropriate treatment methods for the comorbidities [4]. Type 2 is the most common diabetes type with 90% of incidence in the world population [8]. Therefore, we focused to develop our model for Type 2 diabetic patients.

The majority of Machine Learning (ML)-based Type 2 management studies in the literature are related to the early diagnosis of Type 2 Diabetes, and it is noteworthy that there has not been a sufficient amount of work on prediction of Type 2 diabetes-related comorbidities [9]. Therefore, studies to predict the risk of developing comorbidities in Type 2 diabetes patients and to manage this health condition have great importance. This makes it more critical for Type 2 diabetic patients with hypertension because Type 2 diabetes combined with hypertension increases the likelihood of severe comorbidities. This study aims to predict the risk of developing hypertension and reduce the risk of developing further critical comorbidities in Type 2 patients to support lifelong Type 2 diabetes management.

2. Methods

2.1. Data Preprocessing

One of the most important elements in ML-based Type 2 diabetes problems is the real-life representation of the data used. In this study, Primary Care data of patients in the Connected Bradford dataset were used. Connected Bradford Primary Care is a large dataset containing all the observation data of primary care healthcare institutions in Bradford, UK. In this dataset, patients were assigned an anonymous unique ID, and each of the observations resulting from their visits to these primary care health institutions was recorded. The dataset used in this study consists of 1,058,139 patient entries (rows) and each entry has feature (variable) columns such as person id, observation definition, numeric laboratory result, observation date, etc.

However, since this study is a Type 2-related study, these data were filtered only with patients who already had Type 2 diabetes. After this filtering process, a total of over 476,000 Type 2 diabetic patients with over 14,000 variable columns remained. Since this dataset is still very large and calculations are time-consuming, a randomly generated subset of one million rows with 43,000 patients was used. In parallel, the observation definition column was grouped by diseases, and it has been observed that the most common observation is hypertension. To this end, hypertension has been identified as a marker of the risk of Type 2 diabetes progression.

In this study, in accordance with the National Institute for Health and Care Excellence (NICE), patients with 7-day mean systolic and diastolic blood pressure readings of more than 140 mmHg and 90 mmHg, respectively, were categorised as at high risk of developing hypertension [4]. The patients with lower than these reading values were categorised as having a low risk of developing hypertension. Since the finalised data frame is still very big, the columns with no data have been eliminated, and the most 20 frequent variables in this database have been selected to ensure computation and time efficiency. Since there are still some missing values in the data, these missing values were imputed with the average value for each column. Then the resulting new data frame is scaled.

2.2. ML Methods

The most common ML algorithms for Type 2 diabetes-related problems were used, namely [10]: Naïve Bayes (NB), Neural Network (NN), Random Forest (RF), and Support Vector Machine (SVM). These four ML algorithms were trained separately. 80% of the data have been trained using k-Fold cross-validation and parameter hyper-tuning. Cross-validation was used to reduce overfitting and prevent bias. Parameter hyper-tuning was used to optimise the performance of trained models. The remaining 20% of the data was used for testing. Next, Generalized Linear Model (GLM) was used as an ensemble method to combine the predictions and increase the prediction performance of our classification problem.

3. Results

Table 1 shows the Accuracy and Kappa values of the ML methods used. These values are the default metrics used to evaluate algorithms on binary and multi-class classification datasets with the Caret package in R. Accuracy is the percentage of correctly classified instances out of all instances [11]. Kappa or Cohen's Kappa is a classification accuracy, except that it is normalised at the baseline of random chance on the dataset [11]. It is a more useful measure for problems that have an imbalance in the classes [12].

Table 1. Performance values of each ML method and ensemble method

ML Method	Accuracy	Kappa
Naïve Bayes	0.8667	0.1296
Neural Network	0.9500	-0.0034
Random Forest	0.9504	0.0000
Support Vector Machine	0.9500	-0.0008
Ensemble	0.9525	0.2183

Figure 1 shows the importance of each variable in the preprocessed dataset. These importance values have been calculated for the ensemble method using the Caret package in R. In this subset, Body Mass Index – Observation (i.e. body mass divided by the square of the body height), Neutrophil Count (i.e. subset of white blood cells in the immune system), and Serum Cholesterol Level (i.e. combined amount of bad and good cholesterol in blood), are the three most important variables in predicting the risk of hypertension in a patient with Type 2, respectively.

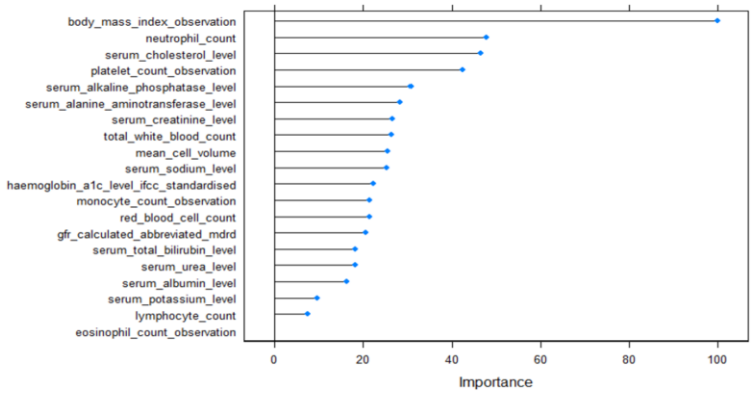

Figure 1. Importance level of each feature in the ensemble method

4. Discussion

In Table 1, RF has the highest accuracy in predicting the risk of hypertension, but the Kappa value, another important performance criterion in classification problems, is 0.0000. Despite its lower accuracy, NB has the highest Kappa value among the four ML algorithms with 0.1296. This shows that accuracy and kappa may give different results with repeated use of ML algorithms, and the performance of the model may vary according to the intended performance criteria. However, when the accuracy and kappa values of the ensemble method are considered, it is seen that both performance criteria have the highest values with 0.9525 and 0.2183, respectively. Even though the accuracy does not increase significantly, the increase in the Kappa value is noteworthy.

In Figure 1, it is observed that the prediction of risk of hypertension in patients with Type 2 diabetes may depend on various variables. However, it has been observed that the importance of variables may vary and especially some variables may have greater importance in the ML model. This causes ML-based prediction outcomes to be more affected by some variables. Seeing the importance levels of the variables in the ML-based models provides an opportunity to make interpretations of the reasons behind the outcomes of the ML model.

5. Conclusion

Hypertension is the most common comorbidity among Type 2 diabetic patients in the Connected Bradford data. In addition, hypertension is related to many serious diseases and has a great significance in predicting the clinically important poor outcomes in Type 2 diabetes, such as heart attack, blindness, or neurological issues. It is important to predict and manage hypertension, which is one of the most important risk factors for developing serious comorbidities in Type 2 diabetic patients. We showed that when high-performing ML models are ensembled, the performance values of the final prediction can potentially increase. However, it is noteworthy to criticise that using different ML algorithms with large datasets has limitations in prediction of Type 2 diabetes progression. Since there are no distinct ML-based methods in progression of Type 2 diabetes, it is crucial to choose the most related data pre-processing techniques and ML

algorithms. It also is useful to provide feature importance of the trained ML model to see the impact of each variable on prediction of Type 2 diabetes progression. Finally, we aim to continue to improve the performance and robustness of the ML algorithms and importantly develop a clinical safety case that considers the assurance of the algorithms in the intended clinical workflow and setting [13]. Beyond performance and safety, we plan to consider the wider ethical issues with the deployment of this type of clinical ML algorithms and explore questions of legal liability and moral responsibility for the outcome of the ML-based decision support systems [14].

Acknowledgment

This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York and by the Engineering and Physical Sciences Research Council (EP/W011239/1).

References

[1] Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. BMJ open. 2013 Jan 1;3(5):e002457. doi: 10.1136/bmjopen-2012002457

[2] K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. InWorkshops at the Twenty-Eighth AAAI conference on artificial intelligence 2014 Jun 18.

[3] Peddinti G, Cobb J, Yengo L, Froguel P, Kravić J, Balkau B, Tuomi T, Aittokallio T, Groop L. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. Diabetologia. 2017 Sep;60(9):1740-50. doi: 10.1007/s00125-017-4325-0

[4] Alonso-Morán E, Orueta JF, Esteban JI, Axpe JM, González M, Polanco NT, Loiola PE, Gaztambide S, Nuño-Solinis R. The prevalence of diabetes-related complications and multimorbidity in the population with type 2 diabetes mellitus in the Basque Country. BMC public health. 2014 Dec;14(1):1-9.

[5] Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports. 2020 Jul 20;10(1):1-2.

[6] Long AN, Dagogo-Jack S. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. The journal of clinical hypertension. 2011 Apr;13(4):244-51.

[7] Ozougwu JC, Obimba KC, Belonwu CD, Unakalamba CB. The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. J Physiol Pathophysiol. 2013 Sep 30;4(4):46-57.

[8] DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI, Simonson DC. Type 2 diabetes mellitus. Nature reviews Disease primers. 2015 Jul 23;1(1):1-22.

[9] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017 Jan 1;15:104-16. doi: 10.1016/j.csbj.2016.12.005

[10] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. Journal of medical Internet research. 2018 May 30;20(5):e10775. doi: 10.2196/10775.

[11] Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008 Nov 10;28:1-26.

[12] Foody GM. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. Remote Sensing of Environment. 2020 Mar 15;239:111630.

[13] Hawkins R, Paterson C, Picardi C, Jia Y, Calinescu R, Habli I. Guidance on the assurance of machine learning in autonomous systems (AMLAS). arXiv preprint arXiv:2102.01564. 2021 Feb 2.

[14] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Apr 4;98(4):251.